Towards an environment for the production and the validation of lexical semantic resources

Mikaël Morardo, Éric Villemonte de La Clergerie

Towards an environment for the production and the validation of lexical semantic resources

Mikaël Morardo, Éric Villemonte de la Clergerie
INRIA-Rocquencourt
Domaine de Voluceau Rocquencourt, B.P. 105, 78153 Le Chesnay, France
{mickael.morardo,eric.de_la_clergerie}@inria.fr

Abstract

We present the components of a processing chain for the creation, visualization, and validation of lexical resources (formed of terms and relations between terms). The core of the chain is a component for building lexical networks relying on Harris' distributional hypothesis applied on the syntactic dependencies produced by the French parser FRMG on large corpora. Another important aspect concerns the use of an online interface for the visualization and collaborative validation of the resulting resources.

Keywords: Terminology extraction, word clustering, visualization interface, collaborative interface, knowledge acquisition

1. Introduction

Each specialized domain tends to have its own set of concepts, instantiated by specialized terms represented by simple or multi-words expressions. Discovering these terms and their relationships is an important issue for providing useful lexical semantic resources (or lexicalized ontologies) for many NLP-based tasks (such as query expansion for search engines, semantic annotation of documents, question answering, translation, ...).

However, hand-crafting such resources remains a fastidious task, which has to be replicated for many domains, and the resources have to be regularly updated, to follow the evolution of a domain (in particular with the emergence of new terms). On the other hand, (unsupervised) acquisition tools are now able to extract automatically many interesting pieces of information from linguistically processed corpora. Unfortunately, these tools still make many errors and often miss important relations (suffering from weak recall). Our opinion is that human validation remains a necessary complement of automatic acquisition, but should be applied on rich data trough well conceived interfaces. Moreover, given the amount of data that has often to be validated, we advocate for collaborative interfaces. These motivations led us to develop a process flow that includes:

1. the deep linguistic processing of corpora (ranging from medium to large sized ones, specialized or not);
2. the extraction of (multi-word) terms and the discovery of semantic proximity between these terms (and simple words), expressed as semantic relations;
3. the visualization and validation of the resulting terms and relations through a collaborative online interface.

The paper is organized as follows: Section 2. introduces some of the corpora we used for our experiments. Section 3. provides some background information about the way the corpora are linguistically processed, in particular to get syntactic data following the PASSAGE annotation scheme. These data are then used for extracting multi-word terms (Section 4.) and for identifying semantically close terms (Section 5.). Finally, the main aspects of the visualization and validation interface are sketched in Section 6..

2. The corpora

As illustrated by the non-exhaustive list of Table 1, we have run our experiments on a large set of French corpora, covering various styles and domains, and with sizes ranging from around one million words to several hundred millions words. The top corpora were prepared in view of the PASSAGE evaluation campaign and constitute the **CPL** (*Corpus Passage Long*) corpus. These corpora have been completed with AFP news to form the **ALL** collection. The **ALL** collection covers various styles (journalistic, encyclopedic, ...) but is not domain specific. The idea is to observe what can be extracted from large non thematic corpora.

On the other hand, the 4 bottom corpora are homogeneous in terms of style and fall in the law domain, covering several more specific subfields (fiscal law, social law, business law, and civil law). These law corpora have been provided by a commercial publisher that wishes to complete and maintain accurate terminology for indexing and querying its collections.

Corpus	#Msent.	#Mwords	Description
Wikipedia (fr)	18.0	178.9	encyclopedic pages
Wikisource (fr)	4.4	64.0	literacy texts
EstRepublicain	10.5	144.9	journalistic
JRC	3.5	66.5	European directives
EP	1.6	41.5	parliamentary debates
Total **CPL**	38.0	495.8	all above
AFP	14.0	248.3	news
Total **ALL**	52.0	744.2	CPL+AFP
fiscal	7.2	145.2	law
social	6.8	127.5	law
civil	2.6	40.9	law
business	7.2	133.8	law

Table 1: Some of the corpora used for the experiments

3. Linguistic processing

All corpora have been processed by the Alpage processing chain[1], with SXPIPE (Sagot and Boullier, 2008) used

[1]freely available at `https://www.rocq.inria.fr/alpage-wiki/tiki-index.php?page=alpc&bl=y`.

for segmentation and named entity recognition (NER), and FRMG used for parsing.

The parser is based on a wide-coverage French Tree Adjoining Grammar (Villemonte de la Clergerie, 2010). The native dependency output of FRMG is converted to the EASy/PASSAGE annotation schema (Vilnat et al., 2010), designed during the two first parsing French evaluation campaigns (EASy and PASSAGE). The PASSAGE scheme is based on a set of 6 kinds of non-recursive chunks and a set of 14 kinds of relations, as described by Table 2. The relations can connect either chunks or forms, and all of them are binary, but for the COORD relations. Figure 1 shows an example of English sentence annotated following the PASSAGE scheme.

Being less rich than FRMG native schema, some information is lost during the conversion to PASSAGE schema. However, the advantage of PASSAGE is to act as some kind of standard, with around 10 parsing systems able to produce it for French. Figure 2 and Figure 3 provide some information about the performances of FRMG on chunks and relations. They have been calculated in 2011 (around the date of our first experiments on the **ALL** corpus) and, more recently, at the end of 2013, on the EasyDev corpus, a small development set of around 4k sentences covering various styles (journalistic, literacy, medical, mail, speech, ...). The improvements between 2011 and 2013 come from a better coverage of FRMG grammar and of the use of training techniques on a treebank for better disambiguation (Villemonte De La Clergerie, 2013).

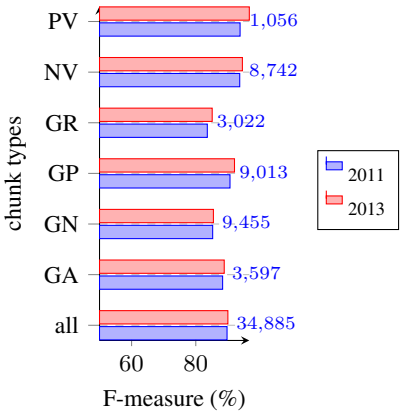

Figure 2: F-measures for Passage chunks (on EasyDev)

From the syntactic results, we collect and count recurring elements of information using a MapReduce algorithm (Dean and Ghemawat, 2004). These elements are then used by the knowledge acquisition scripts presented in the following two sections.

4. Terminology extraction

The first acquisition task concerns the extraction of terms. Terminology extraction still raises some problems but the main ideas are nowadays relatively well identified (Pazienza et al., 2005), in particular for terms corresponding to multi-word expressions. In our experiments, we have focused our work on the extraction of nominal multiword terms that are essentially instances of the pattern

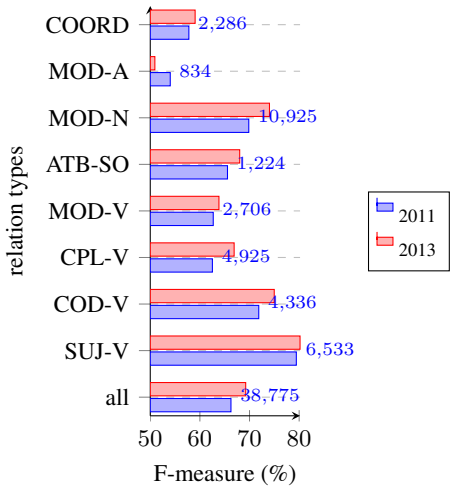

Figure 3: F-measures for Passage relations (on EasyDev)

`(GN)(GR*GA|GP|PV|NV)+` over PASSAGE chunks. This pattern captures nominal chunks [GN] modified by adjectival chunks [GA], prepositional chunks [GP] possibly introducing verbs [PV] or participial verbs [NV], and possibly with some adverbs [GR]. The chunks composing a candidate term must also be syntactically connected (essentially through noun-modifier MOD-N relations). Table 3 show some instances of the pattern for a few terms found in `ALL` corpus.

The candidate terms are then ranked along several criteria, including standard ones such as frequency, internal cohesion (computed via a variation of point-wise mutual information), and more original ones such as autonomy and diversity of contexts.

Autonomy exploits the syntactic dependencies to check that a significant amount of the occurrences of the candidate corresponds to "*active*" syntactic roles (such as subject or object, for instance), and that not all the occurrences are modified (for instance by prepositional chunks). The motivation for the autonomy criterion is to avoid the selection of candidates which are essentially fragments of larger expressions or which play, for instance, the role of adverbial locutions or complex prepositions.

Favoring diversity, we penalize candidates that tend to occur in very similar sentences (or sentence fragments) and are more representative of collocations.[2] Variants are then grouped in function of their underlying lemmas, and some candidates are rejected if their variability is too high, for instance when they include a `_NUMBER`, `_DATE`, or `_LOCATION` lemma that get instantiated by many different named entities[3]).

With minimal filtering (to favor recall), we get around 100K terms on the `all` corpus and around 50K terms on the `fiscal` part of the law corpus (145Mwords). The terms are enriched with a set of randomly chosen illustrative sentences and statistical information. Figure 6 lists some of the terms extracted from the `business` law corpus, with

[2]Favoring diversity is also a way to correct some problems related to duplicated or close sentences, a relatively frequent phenomena in AFP news but also in the other corpora.

[3]but please note that we accept terms built on named entities.

Type	Description
GN	Nominal chunk
NV	Verbal kernel
GA	Adjectival chunk
GR	Adverbial chunk
GP	Prepositional chunk
PV	Prepositional chunk on non-tensed verbal kernel

(a) Chunks

Type	Description
SUJ-V	Subject-verb dep.
AUX-V	Aux-verb dep.
COD-V	direct objects
CPL-V	other verb arguments & complements
MOD-V	verb modifiers (such as adverbs)
COMP	subordinate sentences
ATB-SO	verb attribute
MOD-N	noun modifier
MOD-A	adjective modifier
MOD-R	adverb modifier
MOD-P	prep. modifier
COORD	coordination
APPOS	apposition
JUXT	juxtaposition

(b) Relations

Table 2: PASSAGE annotation scheme

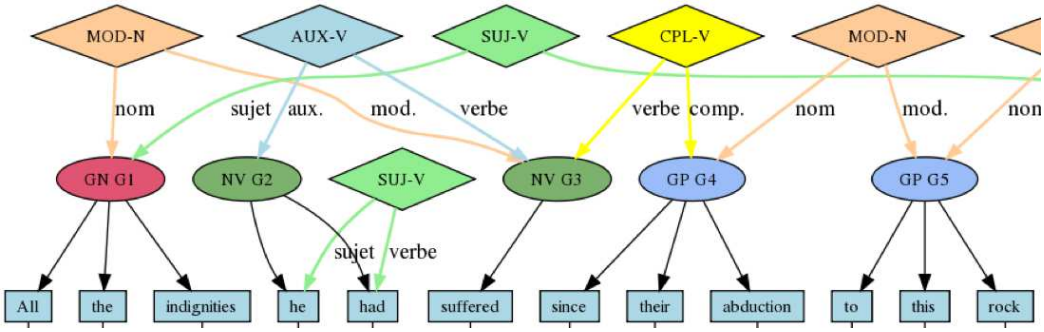

Figure 1: An exemple of English sentence annotated following PASSAGE schema

a focus on *président du conseil* / *Chairman of the Board*. It may be noted that, for *président du conseil*, we observe that several variants of this term have been identified in the corpora, corresponding to several plurals (on *chairman* and *board*) and gender (*chairman*, *chairwoman*).

We are fully aware that terms do not necessarily correspond to multi-word expressions, but we expect the other simple-word terms to be captured when looking for semantic similarity (Section 5.). However, we still need to setup a filtering of the terms to favor domain-specific ones, possibly by contrasting their frequencies with frequencies computed on a reference corpus.

5. Discovering semantic similarities

Most works on semantic clustering (Cimiano et al., 2004; Pantel, 2003) have been inspired by Harris' distributional hypothesis (Harris, 1968) that states that words close semantically tend to occur in similar contexts. Several kinds of contexts have been considered, including bag of words, sliding windows, or, in our case, syntactic contexts derived from syntactic dependencies. For instance, for a CPL-V (complement-verb) dependency triple like $\langle$*to_sit on chair*$\rangle$, one may associate the syntactic context $\langle$*to_sit on* $\bullet\rangle$ to the word *chair* and, in a dual way, the context $\langle\bullet$ *on chair*$\rangle$ to the word *to_sit*. A weighted vector of such contexts may be attached to each word, with the weights reflecting the frequency and importance of the context (measured via mutual information). Table 4 lists the number of occurrences for a few dependency triples involving *chaise* (*chair*). We observe a few actions related to the use of a chair (*assoir sur chaise* and *se assoir sur chaise* [*to sit on a chair*], *tomber sur chaise* [*to fall on a chair*], or *prendre une chaise* [*to take a chair*]), but also many entries corresponding to multi-word terms built upon *chair* (*chaise musical* [*musical chair*] or *chaise électrique* [*eletric chair*]). Obviously, not all high-frequency dependencies are pertinent to capture the meaning of a chair. We can also observe the high frequency of the coordination between *chair* and *table*. For dependencies involving a preposition, we keep triples with the preposition used a relation label. Moreover, we refine the relation label with suffix = when the preposition introduces a noun with no determiner (like *chaise à porteur*).

To counter-balance attachment ambiguity for prepositional groups, we decided to add extra dependencies for potential attachments that were discarded but could have been chosen: for instance, in an expression like *tremblement de terre de magnitude 5* (*earthquake of magnitude 5*), maybe the attachment of *magnitude* was done on *terre* giving triple $\langle terre\ de\ magnitude\rangle$ but we also add the potential attachment $\langle tremblement\ de\ *\ magnitude\rangle$. A similar treatment is done to attach potential dependency triples for the occurrences of candidate (multi-word) terms that may be retrieved in the corpus.

In order to reflect deeper semantic relationships, some of

dioxyde de carbone	*carbon dioxid*	[dioxyde/nc]$_{\text{GN}}$ [de/prep carbone/nc]$_{\text{GP}}$
hockey sur glace	*ice hockey*	[hockey/nc]$_{\text{GN}}$ [sur/prep glace/nc]$_{\text{GP}}$
téléphone portable	*mobile phone*	[téléphone/nc]$_{\text{GN}}$ [portable/adj]$_{\text{GA}}$
lait écrémé	*skimmed milk*	[lait/nc]$_{\text{GN}}$ [écrémer/v]$_{\text{NV}}$
permis de conduire	*driving license*	[permis/nc]$_{\text{GN}}$ [de/prep conduire/v]$_{\text{PV}}$
procréation médicalement assistée	*medically assisted procreation*	[procréation/nc]$_{\text{GN}}$ [médicalement/adv]$_{\text{GR}}$ [assisté/adj]$_{\text{GA}}$
implant chirurgical non actif	*non active chirurgical implant*	[implant/nc]$_{\text{GN}}$ [chirurgical/adj]$_{\text{GA}}$ [non/adv]$_{\text{GR}}$ [actif/adj]$_{\text{GA}}$

Table 3: Examples of terms with their chunk structure

governor	relation	governee	freq.
chaise_nc	et	table_nc	**235**
asseoir_v	sur	chaise_nc	**227**
chaise_nc	modifier	long_adj	**168**
chaise_nc	de=	poste_nc	**115**
tomber_v	sur	chaise_nc	**103**
chaise_nc	modifier	musical_adj	**102**
se_asseoir_v	sur	chaise_nc	**93**

governor	relation	governee	freq.
prendre_v	object	chaise_nc	**87**
chaise_nc	modifier	électrique_adj	**82**
chaise_nc	modifier	vide_adj	**80**
chaise_nc	à=	porteur_nc	**80**
dossier_nc	de	chaise_nc	**78**
avoir_v	object	chaise_nc	**71**
table_nc	et	chaise_nc	**62**

Table 4: A few syntactic dependencies involving *chaise* (*chair*).

the PASSAGE dependencies are rewritten, for instance for passive verbs with the surface subjects transformed into deep objects, or for relating a verb attribute to the subject (rather than to the verb). The relations involving a coordination conjunction are distributed along the coordinated elements.

Given context vectors, a wide spectrum of unsupervised learning techniques have been proposed to regroup words, generally into hard clusters (each word belonging to at most one cluster). We favor the search of relations between words rather than hard clustering, believing that the richness of the words (polysemy and sense shift) makes it difficult to capture meaning through strictly delimited clusters.

Our learning algorithm is derived from Markov clustering (van Dongen, 2000), based on the search of nodes that are connected through a dense set of short paths. Our main contribution is to switch to a bipartite graph connecting (simple or multi-words) terms to contexts, as shown in Figure 4, with $wc_{i,a}$ (resp. $cw_{a,i}$) denoting the weight of context c_i for word w_a (resp. of w_a for context c_i).

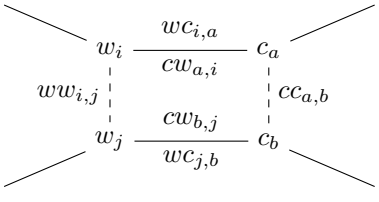

Figure 4: Term-context bipartite graph

The weight $wc_{i,a}$ of context c_a occurring u_{ai} times with word w_i is based on frequency and mutual information, and is given by the following equation, with a similar formulation for the weight $cw_{a,i}$ of w_i relatively to c_a.

$$wc_{i,a} = \frac{ln(u_{ai}) * \eta_a}{\sum_b ln(u_{bi}) * \eta_b} \text{ with } \eta_a = ln\left(\frac{\text{\#distinct words}}{\sqrt{|\{w_j|u_{aj} > 0\}|}}\right) \quad (1)$$

The motivation for a bipartite graph is that terms and syntactic contexts play dual roles: terms sharing similar contexts are semantically close and, conversely, contexts sharing similar terms are also semantically close.

Following (van Dongen, 2000), the search of dense sets of short paths in the graph may be captured by the following set of mutually recursive equations, involving an *inflation* coefficient $\alpha > 1$ than reinforce strong paths over weak ones:

$$\begin{cases} ww_{i,j} = \dfrac{1}{Z_i}\left(\displaystyle\sum_{a,b}(wc_{i,a})(cc_{a,b})(wc_{j,b})\right)^{\alpha} \\ cc_{a,b} = \dfrac{1}{Z_a}\left(\displaystyle\sum_{i,j}(cw_{a,i})(ww_{i,j})(cw_{b,j})\right)^{\alpha} \end{cases} \quad (2)$$

where Z_i and Z_a denote normalization factors given by

$$\begin{cases} Z_i = \displaystyle\sum_{j}\left(\sum_{ab}(wc_{i,a})(cc_{a,b})(wc_{j,b})\right)^{\alpha} \\ Z_a = \displaystyle\sum_{b}\left(\sum_{i,j}(cw_{a,i})(ww_{i,j})(cw_{b,j})\right)^{\alpha} \end{cases} \quad (3)$$

These equations may be reformulated with matrices, using an *inflation* operator Γ_α (with normalization), as follows, with the similarity matrices $W = (ww_{i,j})_{i,j}$, $C = (cc_{a,b})_{a,b}$, and the weight matrices $F = (wc_{i,a})_{i,a}$, $G = (cw_{a,i})_{a,i}$:

$$\begin{cases} W = \Gamma_\alpha(F^t C F) \\ C = \Gamma_\alpha(G^t W G) \end{cases} \quad (4)$$

The formulation involves mutually recursive equations which require the search of a fixpoint, whose solution is approached through an iterative algorithm, starting from initial similarity matrices $W^{(0)}$ and $C^{(0)}$.

The base algorithm is extended by exploiting *transfer matrices* using to transfer the similarities found between words at the level of contexts and conversely. Indeed, the contexts are built upon words (for instance $\langle$*to_sit on* $\bullet\rangle$ is built upon

to_sit by combining it with relation *on*), and one may expect contexts built upon similar words (and same relation r) to be themselves similar. We therefore introduce a transfer coefficient β (set to 0.2 by default) and transfer matrices $T_r = (\tau_{ia})_{i,a}$ for each relation r (such as `object`) with $\tau_{ia} = 1$ if $c_a = r.w_i$ and 0 otherwise. Equations (4) are then modified as follows:

$$\begin{cases} W = \Gamma_\alpha(F^t C F + \sum_r \beta T_r^t C T_r) \\ C = \Gamma_\alpha(G^t W G + \sum_r \beta T_r W T_r^t) \end{cases} \quad (5)$$

The algorithm can also be easily enriched to handle all kinds of extra sources of information about known or assumed similarities between words or contexts. In particular, bonus/malus matrices may be added to provide similarity bonuses or maluses between pairs of words, coming for instance from some external source (like Wordnet (Fellbaum, 1998)). In practice, we add such bonuses for the following cases:

- between w_i and itself, to enforce self-similarity;
- between words that are frequently coordinated (like *chair* and *table*);
- between words close for the editing distance (often reflecting typographic errors or diacritic variations);
- between words sharing common prefixes or suffixes (reflecting some common origin).

More formally, we consider a bonus/malus matrix L added to the identity matrix I, to get the following equation for W, where $\circ$ denotes the point-wise Hadamard product:

$$W = \Gamma_1((I + L) \circ \Gamma_\alpha(F^t C F + \sum_r \beta T_r^t C T_r)) \quad (6)$$

One of the strengths of the algorithm comes from the possibility to retrieve the most pertinent contexts that explain the semantic similarity between two terms w_i and w_j. It may be noted that a term may be related to several other terms through (completely or partially) distinct sets of pertinent contexts, illustrating its polysemy or sense shifts. For instance, from the **ALL** corpus, we found that the words *char* (in the sense of *carriage*) was close of *charrette* (*cart*) and *chariot* (*trolley*) because of contexts like *atteler* (*to harness*) or *promener en X* (*X ride*) while *char* (in the sense of *tank*) was close of *tank* because of contexts like $\langle \bullet$ *de combat*$\rangle$ ($\langle \bullet$ *of combat*$\rangle$) and $\langle$*régiment de* $\bullet\rangle$ ($\langle$*regiment of* $\bullet\rangle$).

These contexts are also useful for an human to assess the validity of the semantic relations.

On the **ALL** corpus (without injecting the extracted terms), the algorithm returned a set of 51,980 pairs (w_i, w_j), involving 19,960 words w_i (including a large number of named entities). By symmetrizing these non-necessarily symmetric pairs, we obtain a large non-oriented network with 47,065 edges. For the `busyness` corpus (with extracted terms included), we get a non-oriented network with 10,223 nodes and 13,584 edges.

Figure 5 shows a tiny part of the **ALL** network, centered on *jambe* (*leg*) and displayed with Tulip software[4] (Auber, 2003). We clearly observe a bush-like structure, with a set of bodypart terms strongly interconnected that form a good cluster, more precisely related to bony and muscular parts. Many other such bush structures were actually identified, which led us to design a small algorithm to extract hard clusters from them, with some of the around 4000 extracted clusters listed below:

79: (a cluster of various kinds of dogs) *sulky malinois fox-terrier setter cocker colley chiot fox labrador ratier griffon caniche teckel épagneul*

80: (a cluster of various kinds of soldiers and military groups) *arrière-garde canonnier cavalerie carabinier tirailleur hussard panzer voltigeur blindé grenadier cuirassier avant-garde zouave lancier*

83: (a cluster of various kinds of diseases) *pneumonie paludisme diphtérie pneumopathie variole dysenterie malaria botulisme poliomyélite septicémie varicelle polio rougeole méningite*

6. Visualization and collaborative validation

Tulip already offers a nice way to view and navigate in the semantic network. However, it is not always adequate for exploring dense areas and is not designed to validate or invalidate relations. Furthermore, it is also not possible to access the explanations motivating a relation, even if they are provided by the acquisition algorithm.

A first step was to complete the subjective intuition provided by Tulip by more objective global evaluations using wordnet-like resources for French as reference resources, for instance by answering automatically and randomly built TOEFL tests (Turney, 2002). Such a test is given by a list of questions, each question specifying a candidate term and a list of 4 potential answer terms, with only one being really close semantically from the candidate term. The success rate when answering randomly is therefore of 25%. We build the tests using two French wordnets, namely French EuroWordnet (Jacquin et al., 2007), and Wolf (Sagot and Fišer, 2008). For each question, the right answer term is selected (randomly) in the same synset than the candidate term. while the other terms are selected (randomly) in other synsets. The results are presented in Table 5. These evaluations essentially provide global information about the recall and precision of the extracted network, and, although the precision may be good (especially for nouns with 94% of good answers, but less for adverbs with only 49%), we mostly observe a weak recall (a low 35% for nouns) as shown in Table 6. We also observed that many relations present in the network but not present in the reference resources may be considered pertinent by an human and it may be noted that comparing two wordnets together (such as Wolf with French EuroWordNet) show that even these reference resources do not provide the same information (with a success rate of 64.5%).

[4]Tulip may be found at `http://tulip.labri.fr/TulipDrupal/` and other examples of visualization of the `all` network with Tulip may be found online at `http://alpage.inria.fr/~clerger/wnet/wnet.html`.

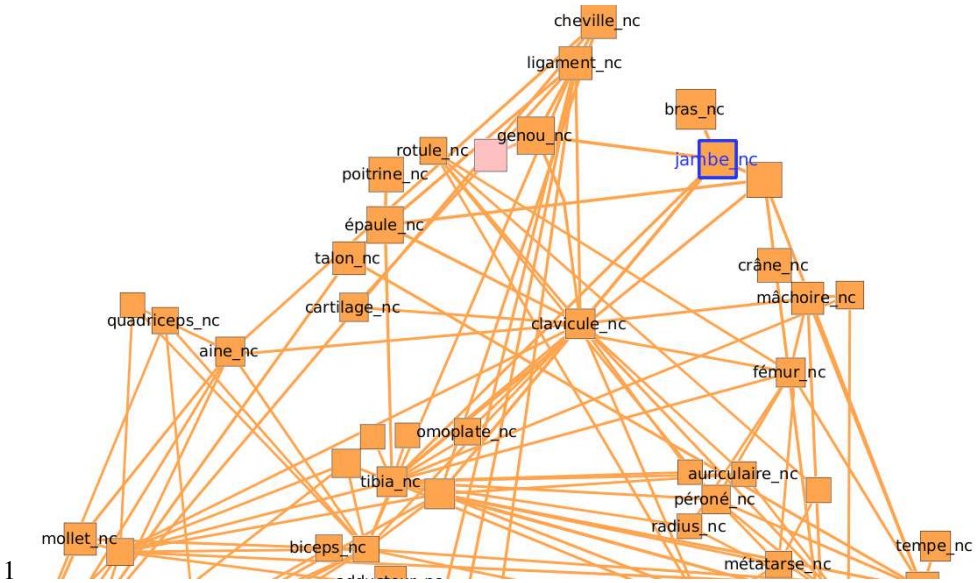

Figure 5: Network fragment, centered on *jambe* (*leg*), mostly listing body parts, viewed with Tulip software

	fwn		wolf	
corpus	**%ok**	**#tests**	**%ok**	**#tests**
all	51,5	4,121	42,1	7,674
fiscal	46,1	104	37,0	493
affaires	35,1	248	43,2	1,055
social	39,4	274	37,7	1,345
wolf	64,5	1,076		

Table 5: Toefl evaluation.

pos	**#tests**	**%ok**	**%bad**	**%missing**	$\%b/(b+f)$
v	3,876	35,5	30,9	33,6	53,4
nc	1,078	33,5	2,1	64,4	94,0
adj	2,085	22,3	11,3	66,4	66,3
adv	1,533	36,9	41,9	21,7	46,8

Table 6: Tests Toefl by syntactic categories (on **CPL**).

Therefore, we finally opted for the development of an online interface[5] for viewing, navigating, and editing the semantic networks and the candidate terms extracted by our acquisition algorithms. Because of the large size of the extracted resources, we also believe that a collaborative approach is needed, hence motivating the choice of an online interface. The implementation was done under the LIBELLEX platform, in the context of a collaboration with Lingua & Machina, the company developing this platform, primarily for the maintenance of multilingual resources for translation.

Figure 6 shows some elements of visualization provided by the interface via several tiles. One of the tile is used to list, query, edit, and validate the terms. For a given term, another tile provides access to illustrative sentences and to statistical explanations. However, the most useful tile (in our opinion) displays a small local graph centered on some selected term (*president of the board* for Figure 6), with the display of the semantic relations but also of structural relations derived from the internal structure of the multi-word terms (such as term expansion or term embedding). Only neighbors up to distance 2 are displayed for clarity using a force directed algorithm, implemented within javascript library `d3.js`. The algorithm tends to nicely separate the clusters (with attractive forces inside the clusters and repulsive ones outside the clusters). In Figure 6, the terms close from *president of the board* include terms related to function or statute, like *vide-président* (*vice-president*), *directeur* (*director*), *administrateur* (*administrator*), or related to membership, like *membres du conseil* or *membre du directoire* (*board members*). However, even if the relations for this example are interesting, it seems necessary to slightly re-organize them and to add a few missing ones, which can done through the interface.

A single glimpse is often enough to quickly detect anomalies and browsing may be done by simply clicking on a node to select it and recenter the graph on it. However, when one need to understand more precisely why several terms are close, it is possible to get more precise information by selecting the associated nodes and opening a new tile that displays a synthetic matrix listing the most pertinent contexts (and their strength) behind the relations for these nodes, as illustrated by Figure 7. These matrices are generally very useful for understanding why terms have been grouped together and are completed by illustrative sentences for the terms and contexts. It is worthwhile to mention that this functionality has proven its usefulness in several occasions where the first intuition of an human was to wrongly discard a relation. Interestingly, for the terms listed in Figure 7 corresponding to a few body parts (*ankle*, *toe*, *wrist*), most of the relating contexts correspond to damages (*fracture*, *sprain*, ...) and pain. Looking at the illustrative sentences, we see that the contexts were actually extracted from journalistic AFP news about sport, which shows how the prox-

[5] accessible at `http://alpage.inria.fr/Lbx` with login `guest` and password `guest`, selecting for instance `allsemnet` under `demo`.

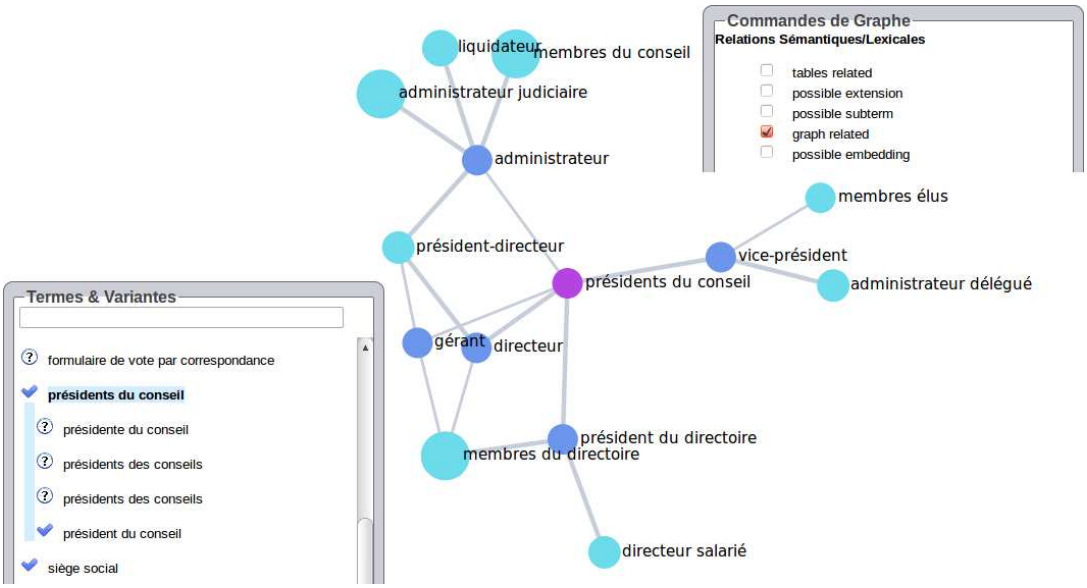

Figure 6: Visualization with Libellex (fragment of Law subcorpus `busyness`), centered on *presidents of the board*

imity between terms is not necessarily intrinsic but also related to some point of view.

7. Conclusion

We propose a complete set of components for the creation, visualization, and collaborative editing of lexical semantic resources.

The linguistic processing chain and the acquisition modules could be easily replaced by similar modules, and the most crucial component is maybe finally the online interface.

In particular, in addition of the extracted terms, the law publisher has also inserted (through merging) a list of potential terms that they have accumulated over the years and that they also wanted to validate (totalling 107K terms for the fiscal part, for instance, to be contrasted with the 50K extracted terms). They routinely use the interface for validating the terms, with around 45K terms accepted for the fiscal part (out of the extracted and added terms). They now plan to explore the validation of the relations in a second stage. Their feedback was helpful to improve the design and the functionalities of the interface and we also expect to exploit the validated data to improve our acquisition algorithms, in particular through the training of a reranker for the terms.

It is also interesting to mention the strong potential of the interface for many similar kinds of lexical semantic resources. In particular, we have loaded WOLF (Sagot and Fišer, 2008), a freely available version of a French Wordnet, with several kinds of lexical relations between synsets. We have also noted, several times and for various audiences (including children), the impact of the graph view for presenting and navigating in rich lexical networks.

Our ambition is now to largely open the service for experiments and feedback with various kinds of lexical semantic resources. Our linguistic processing chain and the acquisition tools are freely available (on the INRIA GForge) but we also plan to offer online processing service for small corpora (up to 1 million words), coupled with the use of the interface.

8. References

Auber, D. (2003). Tulip : A huge graph visualisation framework. In Mutzel, P. and Jünger, M., editors, *Graph Drawing Softwares*, Mathematics and Visualization, pages 105–126. Springer-Verlag.

Cimiano, P., Staab, S., and Hotho., A. (2004). Clustering ontologies from text. In *Proceedings of LREC'04*, pages 1721–1724.

Dean, J. and Ghemawat, S. (2004). MapReduce: Simplified data processing on large clusters. In *OSDI'04: Sixth Symposium on Operating System Design and Implementation*, San Francisco, CA, December.

Fellbaum, C., editor. (1998). *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May.

Harris, Z. (1968). *Mathematical Structures of Languages*. John Wiley & Sons, New-York.

Jacquin, C., Desmontils, E., and Monceaux, L. (2007). French eurowordnet lexical database improvements. In *In Proc. of CICLing'07*, number 4394 in LNCS, Mexico City, Mexico.

Pantel, P. (2003). *Clustering by Committee*. Ph.d. dissertation, Department of Computing Science, University of Alberta, Canada.

Pazienza, M. T., Pennacchiotti, M., and Zanzotto, F. M. (2005). Terminology extraction: an analysis of linguistic and statistical approaches. In (Ed.), S. S., editor, *Knowledge Mining*, volume 185 of *Studies in Fuzziness and Soft Computing*. Springer Verlag.

Sagot, B. and Boullier, P. (2008). SxPipe 2 : architecture pour le traitement présyntaxique de corpus bruts. *Traitement Automatique des Langues (T.A.L.)*, 49(2):155–188.

Sagot, B. and Fišer, D. (2008). Construction d'un wordnet libre du français à partir de ressources multilingues. In *TALN 2008*, Avignon, France.

Turney, P. D. (2002). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. *CoRR*, cs.LG/0212033.

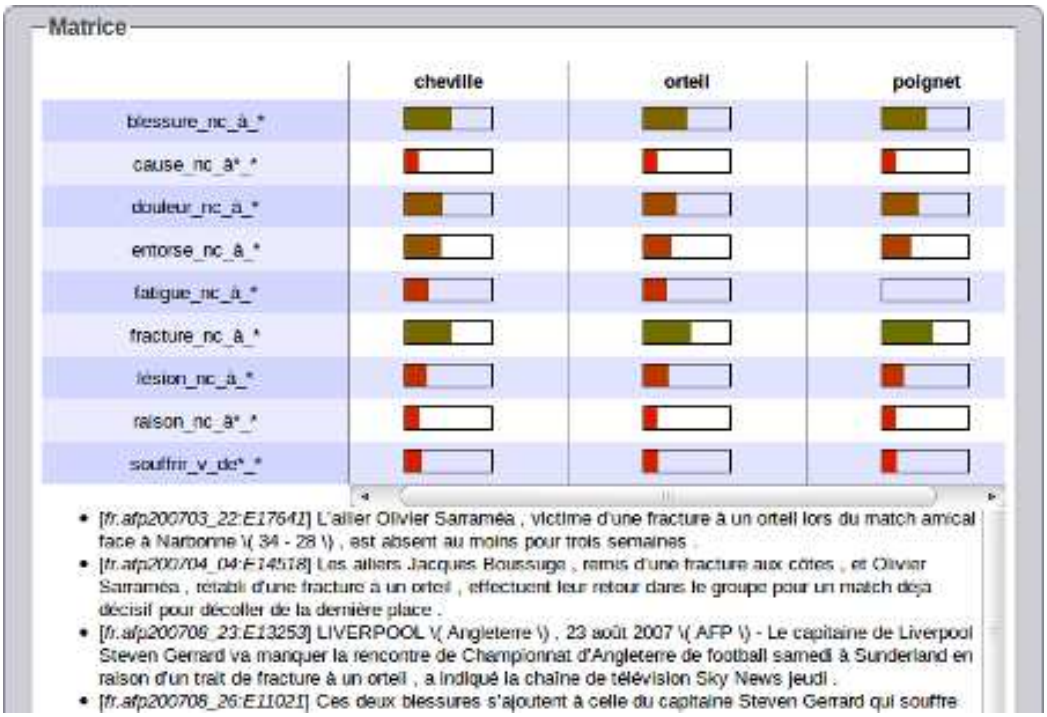

Figure 7: A term-context matrix with illustrative sentences

van Dongen, S. (2000). *Graph Clustering by Flow Simulation*. Phd thesis, University of Utrecht, May.

Villemonte de la Clergerie, E. (2010). Building factorized TAGs with meta-grammars. In *TAG+10: The 10th International Conference on Tree Adjoining Grammars and Related Formalisms*, pages pp. 111–118, New Haven, CO.

Villemonte De La Clergerie, É. (2013). Improving a symbolic parser through partially supervised learning. In *The 13th International Conference on Parsing Technologies (IWPT)*, Nara, Japon.

Vilnat, A., Paroubek, P., Villemonte de la Clergerie, E., Francopoulo, G., and Guénot, M.-L. (2010). PASSAGE syntactic representation: a minimal common ground for evaluation. In *LREC*, La Vallete.